nature biotechnology

Haplotyping germline and cancer genomes with high-throughput linked-read sequencing

Grace X Y Zheng[1,4], Billy T Lau[2,4], Michael Schnall-Levin[1], Mirna Jarosz[1], John M Bell[2], Christopher M Hindson[1], Sofia Kyriazopoulou-Panagiotopoulou[1], Donald A Masquelier[1], Landon Merrill[1], Jessica M Terry[1], Patrice A Mudivarti[1], Paul W Wyatt[1], Rajiv Bharadwaj[1], Anthony J Makarewicz[1], Yuan Li[1], Phillip Belgrader[1], Andrew D Price[1], Adam J Lowe[1], Patrick Marks[1], Gerard M Vurens[1], Paul Hardenbol[1], Luz Montesclaros[1], Melissa Luo[1], Lawrence Greenfield[1], Alexander Wong[1], David E Birch[1], Steven W Short[1], Keith P Bjornson[1], Pranav Patel[1], Erik S Hopmans[2], Christina Wood[3], Sukhvinder Kaur[1], Glenn K Lockwood[1], David Stafford[1], Joshua P Delaney[1], Indira Wu[1], Heather S Ordonez[1], Susan M Grimes[2], Stephanie Greer[3], Josephine Y Lee[1], Kamila Belhocine[1], Kristina M Giorda[1], William H Heaton[1], Geoffrey P McDermott[1], Zachary W Bent[1], Francesca Meschi[1], Nikola O Kondov[1], Ryan Wilson[1], Jorge A Bernate[1], Shawn Gauby[1], Alex Kindwall[1], Clara Bermejo[1], Adrian N Fehr[1], Adrian Chan[1], Serge Saxonov[1], Kevin D Ness[1], Benjamin J Hindson[1] & Hanlee P Ji[2,3]

Haplotyping of human chromosomes is a prerequisite for cataloguing the full repertoire of genetic variation. We present a microfluidics-based, linked-read sequencing technology that can phase and haplotype germline and cancer genomes using nanograms of input DNA. This high-throughput platform prepares barcoded libraries for short-read sequencing and computationally reconstructs long-range haplotype and structural variant information. We generate haplotype blocks in a nuclear trio that are concordant with expected inheritance patterns and phase a set of structural variants. We also resolve the structure of the *EML4-ALK* gene fusion in the NCI-H2228 cancer cell line using phased exome sequencing. Finally, we assign genetic aberrations to specific megabase-scale haplotypes generated from whole-genome sequencing of a primary colorectal adenocarcinoma. This approach resolves haplotype information using up to 100 times less genomic DNA than some methods and enables the accurate detection of structural variants.

The human genome is diploid, with each cell containing a copy of both the maternal and paternal chromosomes. A comprehensive understanding of human genetic variation requires identification of the order, structure and origin of these sets of alleles and their variants across the genome[1]. Haplotypes, the contiguous phased blocks of genomic variants specific to a given homolog, are essential to such analysis. Genome-scale haplotype analysis has many advantages in genetic studies. Phasing of germline variants can be used to identify causative mutations in pedigrees, determine the structure of genomic rearrangement events and unravel *cis*- versus *trans*-relationships of ostensibly linked variants. In the case of cancer genomes, phasing studies provide insight into the haplotype context of somatic mutations, genomic rearrangements critical for oncogenesis and aneuploidy of chromosomes[2].

Traditionally, haplotype analysis has relied on statistical inference from the genomic sequences of related individuals[3]. More recently, experimental sequencing approaches have been used[1,2,4–9]. Next-generation sequencing (NGS) methods for determining haplotype require either subcloning into fosmids[1,2] or sample dilution into 96- or 384-well microtiter plates for sequencing library preparation[6–9]. In dilution haplotyping, the low molarity of molecules per partition reduces the likelihood that one DNA molecule has overlapping sequence with another—and, thus, that haplotypes can be derived. However, current dilution methods require complicated experimental protocols, high amounts of starting DNA, extensive manipulation of DNA and microwell plates with restricted partitioning capacity[6–9]. Recently, experimentally phased genome assemblies have been demonstrated with long-read technology[10], but high amounts of input DNA requirements, lower sequence read accuracy and cost create substantial barriers for its widespread adoption.

Here we describe a microfluidic technology that circumvents these complex experimental protocols and generates haplotype-resolved genome sequences using small amounts of input DNA. Approximately 300 genomic equivalents, or 1 ng of high–molecular weight (HMW) genomic DNA, is distributed across more than 100,000 droplet partitions, where it is barcoded and subject to random priming and

[1]10X Genomics, Pleasanton, California, USA. [2]Stanford Genome Technology Center, Stanford University, Palo Alto, California, USA. [3]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to H.P.J. (genomics_ji@stanford.edu) or B.J.H. (ben@10xgenomics.com).

Received 16 May 2015; accepted 12 November 2015; published online 1 February 2016; doi:10.1038/nbt.3432

polymerase amplification. Barcode-tagged DNA molecules are released from each droplet, after which a modified library preparation is done. The resulting libraries then undergo standard Illumina short-read sequencing. A computational algorithm uses the barcodes to link sequencing reads to the originating HMW DNA molecule, enabling the construction of contiguous segments of phased variants. Loading nanogram amounts of DNA into large number of gel beads minimizes the chance of coincidental barcode overlap and improves overall phasing performance. Furthermore, the higher number of barcodes per amount of genomic DNA improves detection of structural variants. Using exome enrichment, we determine gene-level phasing in a nuclear family trio and define the complex structure of an oncogenic rearrangement in a cancer cell line. Finally, we generate a phased genomic analysis of a primary colorectal cancer originating from a clinical tissue sample.

RESULTS

Microfluidics and gel bead partitioning

Our high-throughput, droplet-based reagent delivery system uses hydrogel beads (gel beads) to deliver barcoded oligonucleotides. Within the confines of a microfluidic chip, monodispersed gel beads are deformable and can be delivered with a fill rate of more than 85% per droplet[11]. An individual gel bead is functionalized with millions of copies of the same barcoded oligonucleotide. Droplet partitions number more than 100,000, and this feature exceeds what can be achieved with microwell systems or virtual partitioning[9] by orders of magnitude.

Reagent delivery and sample partitioning are performed in a plastic microfluidic consumable cartridge that processes eight samples simultaneously. Cartridge reservoirs are loaded with gel beads, the sample and reagent mixture and an oil-surfactant solution. Reagents are delivered from the reservoirs via a network of microfluidic channels to a microfluidic 'double-cross' junction (**Fig. 1a**). The first junction combines a close-packed aqueous slurry of gel beads with the sample and reagent mixture, and the second junction delivers the oil-surfactant solution. Droplet generation occurs at a rate of ~1 kHz at the second junction, resulting in more than 100,000 droplets loaded with >85% single occupancy per droplet partition. The droplets flow to a collection reservoir, where they are then transferred to a conventional 96-well plate. A chemical reducing agent in the reaction mix dissolves the gel beads, triggering the release of the barcoded oligonucleotides from the gel matrix. As a result, independent solution-phase reactions are conducted without further addition of reagents.

In our study, 300 genomic equivalents (1 ng) are dispersed into >100,000 droplet partitions with different barcodes; thus, only a small number of genomic equivalents are loaded per partition. We found that it is highly unlikely that two distinct HMW molecules that cover the same loci but have opposing haplotypes will a barcode ($P < 0.002$). The use of fewer barcodes would require even smaller amounts of input DNA—equivalent to only a few genome equivalents—to maintain low rates of intermolecular genomic overlap within the same barcoded partition. Such low amounts of input DNA would lead to problems with allelic dropout and coverage gaps throughout the genome.

Barcode sequencing

To create barcoded DNA molecules for sequencing, we perform an optimized droplet-based assay that introduces a barcode-containing sequencing adapter into new fragments (Online Methods). HMW DNA templates, ranging from ten to several hundred kilobases in size, are randomly distributed in picoliter reaction volumes across >100,000 droplets. Within an individual droplet, gel bead dissolution releases the amplification primer into the partitioned solution. The primer contains the following components: (i) an Illumina P5 flow cell primer sequence, (ii) a 14-bp barcode, (iii) an Illumina R1 sequence (read 1 sequencing primer) and (iv) a 10-bp random primer sequence (**Supplementary Fig. 1**). Amplification is nonprocessive, and the length of molecules produced range from a few to several hundred kilobases. After thermocycling of the droplets, the emulsion is broken, the pooled aqueous fractions are recovered, and the library preparation is completed with the addition of other adapter components (additional details of library preparation and metrics of DNA loading are noted in Online Methods).

Whole-genome sequencing with linked reads

To assess the performance of linked-read sequencing for haplotyping, we relied on three HapMap samples that form a nuclear trio: NA12878 (mother), NA12877 (father) and NA12882 (child). These samples have been experimentally phased and have haplotypes statistically derived from inheritance patterns[9,12].

We generated barcode sequencing libraries from this trio. Approximately 1 ng of HMW DNA from each sample was processed on a 10X GemCode instrument, and sequencing libraries were prepared. The barcode libraries underwent whole-genome sequencing (WGS) to ~30× mean coverage. We used an Illumina Hiseq 2500 with 2 × 98 paired-end reads (**Supplementary Table 1**) for all sequencing analyses reported here. Reads were trimmed of the first 10 bp to remove primer extension artifacts and aligned to the human reference genome (hg19) with BWA[13], and PCR duplicates were marked using

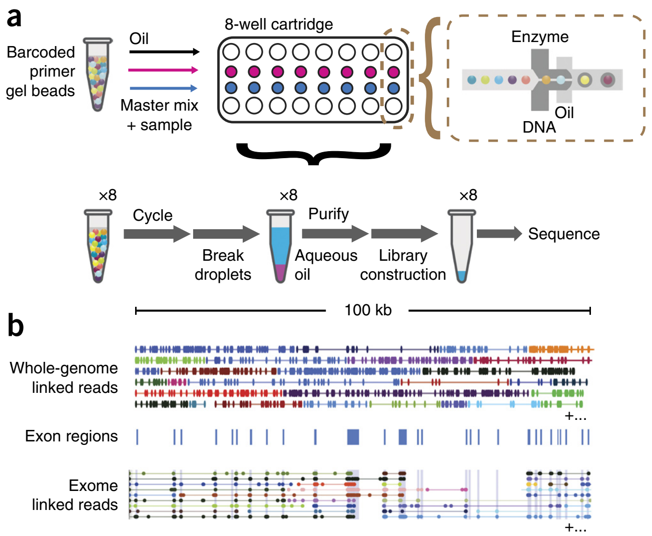


Figure 1 Overview of the technology for generating linked reads. (**a**) Gel beads loaded with primers and barcoded oligonucleotides are mixed with DNA and enzyme mixture then oil-surfactant solution at a microfluidic 'double-cross' junction. Gel bead–containing droplets flow to a reservoir where gel beads are dissolved, initiating whole-genome primer extension. The products are pooled from each droplet. The final library preparation requires shearing the libraries and incorporation of Illumina adapters. (**b**) Top, linked reads of the *ALK* gene from the NA12878 WGS sample. Lines represent linked reads; dots represent reads; color indicates barcode. Middle, exon boundaries of the *ALK* gene. Bottom, linked reads of the *ALK* gene from the NA12878 exome data. Reads from neighboring exons are linked by common barcodes. Only a small fraction of linked reads is presented here.

Table 1 Summary of phasing results

Phased WGS	NA12878[a]	NA12877[a]	NA12882[a]	NA20847[b]	Patient 1532 normal[c]	Patient 1532 tumor[c]
Fold coverage	37	34	36	32	31	32
% aligned	97	97	96	95	96	96
% duplication	1.38	1.38	1.19	6.45	0.5	0.7
Relative genomic equivalents per partition	0.0015	0.002	0.0019	0.0044	0.0015	0.0011
% barcodes correctly assigned	92	87	85	91	88	86
Effective barcode diversity	111,808	151,552	126,326	144,442	136,888	125,694
Number of molecules (millions)	15	34.5	23.4	43.7	30	20.6
Length-weighted mean molecule length (kb)	60.53	40.20	47.35	75.69	44.35	38.31
% SNPs phased	99	97	99	98	95	95
% genes phased (<100 kb)	97	92	97	95	93	91
N50 phase block (kb)	2,834.44	890.08	1,726.40	2,577.04	1,570.40	962.11
Longest phase block (kb)	14,557.82	7,016.43	11,545.49	12,729.94	8,729.79	6,706.66
SNV short switch error rate (%)	0.01	0.48	0.20	0.93	N/A	N/A
SNV long switch error rate (%)	0.01	0.03	0.02	0.09	N/A	N/A
Phased exome sequencing	NA12878[a]	NA12877[a]	NA12882[a]	NA20847[b]	NCI-H2228	
Fold coverage	239	497	443	185	244	
% bases on target	57	62	59	62	52	
% aligned	99	99	99	99	99	
% duplication	8.09	18.03	13.79	4.76	8.26	
Relative genomic equivalents per partition	0.0007	0.0010	0.0011	0.0008	0.0007	
% barcodes correctly assigned	92	92	92	87	87	
Effective barcode diversity	149,170	138,086	150,115	154,682	121,906	
% genes phased (<100 kb)	95	95	96	92	N/A	
N50 phase block (kb)	136.54	103.42	113.65	83.94	N/A	
Longest phase block (kb)	2,086.24	1,217.01	1,199.03	957.57	N/A	
SNV short switch error rate (%)	0.47	1.10	0.99	1.36	N/A	
SNV long switch error rate (%)	0.05	0.11	0.07	0.06	N/A	

[a]Ground truth is from ref. 15; [b]ground truth is from ref. 1; [c]ground truth is from Genome Analysis Toolkit call of ~50× WGS of these samples from our analysis. SNP, single nucleotide polymorphism.

the barcode information and alignment position (Online Methods and **Supplementary Fig. 1**). Overall, library quality was exceptionally high; >95% of reads were mapped, and >90% of the genome was covered at least once, with most gaps shorter than 50 kb (**Table 1**, **Supplementary Table 1** and **Supplementary Fig. 2**). PCR duplication rates were minimal, at <1.4% (**Table 1**). Library insert size averaged ~200 bp, and the number of binding events was estimated to be ~45,000 within each partition (**Supplementary Table 2** and Online Methods). Barcodes were uniformly distributed over 100,000 partitions based on an effective number of barcodes adjusted for unevenness of barcode counts (Online Methods, **Table 1** and **Supplementary Fig. 2**). The GC distribution (**Supplementary Fig. 2d**) shows the predictable effect of the nonprocessive amplification step.

Some droplets may coincidentally receive identical barcodes; performance for haplotyping is maintained by loading only a small amount of genomic DNA into each droplet partition. Conversely, a high DNA input amount increases the coincidental loading of overlapping DNA molecules to the same droplet (and, thus, barcode); the overlap among DNA templates obscures haplotype analysis. As a metric to gauge coincidental loading, we calculated relative genome loading per droplet partition. From nuclear trio WGS data, the relative genomic loading was <0.002 genome equivalents per partition (**Table 1**). Thus, such loading issues are minimized given the low molarity distribution of DNA per droplet.

A feature of our analysis involves the concept of linked reads, which are sequences that have the same barcode and are determined to be in physical proximity on the basis of alignment (**Fig. 1b** and Online Methods). This type of data enables the inference of input DNA length, phasing of haplotype blocks and identification of DNA molecules with structural changes (for example, genomic fusion) compared to the reference. The mean number of linked reads per DNA molecule was ~15 for the trio libraries. With linked reads, we determined that these libraries had an inferred length-weighted mean DNA molecule length of at least 40 kb, with a maximum length reaching 200 kb (**Fig. 2a**). This is consistent with the distribution of input DNA detected by gel electrophoresis (**Supplementary Fig. 2b**).

Using low genomic fractions per barcode achieves a high number of linked reads per DNA molecule. We load approximately 3 Mb of genomic DNA per partition. In comparison, whole-genome sequencing by contiguity-preserving transposition (CPT-seq) averages ~21–62 Mb genomic DNA per partition[9]. Phasing of variants becomes substantially more robust; the high number of barcode partitions in combination with the low number of haploid genome equivalents reduces the probability of a partition having two HMW molecules overlapping the same genomic loci but with opposing haplotypes. As a result, the method achieves excellent phasing performance, requires low amounts of genomic DNA and relies on a standard WGS coverage to achieve larger haplotypes, as described below. Further decreasing the amount of genomic DNA per partition will accordingly increase the number of linked reads per molecule. However, the variance in the amount of genomic DNA loaded per partition substantially increases, resulting in allelic dropout and uneven coverage. We have observed that our approach is robust within twofold of the optimal DNA loading amount (data not shown).

Without using standard WGS and after single nucleotide variant (SNV) calling[14], our method can achieve an SNV calling sensitivity (a measure of true positive rate) of 0.93 and a positive predictive value (PPV, a measure of precision) of 0.99 at 30× sequencing of NA12878. In contrast, a TruSeq library at 0.99 PPV has a SNV calling sensitivity of 0.98. We compared the coverage distribution of barcode libraries with Illumina TruSeq libraries (**Supplementary Fig. 3**). Overall, 90% of all genomic bases were covered by both samples. We note that the genome coverage of GemCode libraries was broader and biased against GC-rich regions (**Supplementary Fig. 2d**). However, the metrics observed with standard TruSeq libraries require substantially higher amounts of input DNA; Illumina TruSeq libraries made from

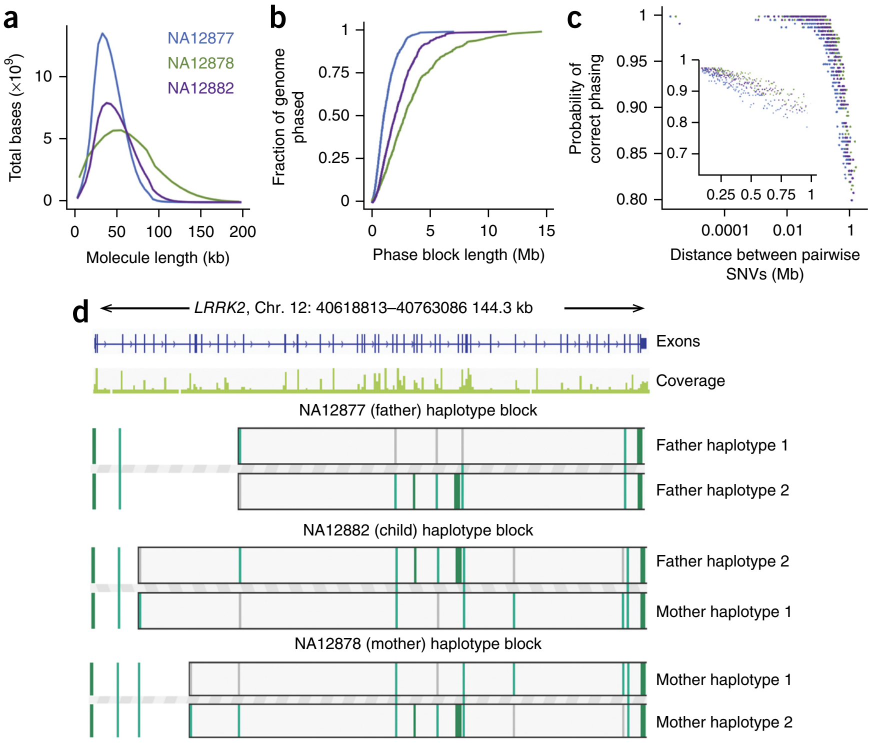


Figure 2 Phasing performance of NA12878 trio analysis. (**a**) Length-weighted molecule size of the trio WGS data, calculated as the number of molecules in the length bin × the median of the length bin. (**b**) Cumulative distribution function of phase block length of the trio WGS samples. (**c**) Phasing accuracy. For all pairs of SNVs that are on the same phasing block, the probability of correct phasing of a pair is plotted as a function of its distance. Inset, SNV pairs separated by at least 0.1 Mb. (**d**) Haplotype blocks of *LRRK2* of the trio exome libraries demonstrating Mendelian inheritance. Most of this gene is phased in all trio samples, but the beginning is not. NA12882 (child) inherited one allele from haplotype 2 from NA12877 (father) and haplotype 1 from NA12878 (mother). Gray bars represent reference alleles; green bars represent alternative alleles.

1 ng of input generally have high proportions of PCR duplicates.

Experimentally derived haplotypes from control genomes

We used linked reads to phase SNVs that were previously annotated (**Table 1** and **Supplementary Table 3**) across all three samples in the nuclear trio. We used a maximum likelihood approach to find near-optimal local haplotype assignments based on read and barcode support (Online Methods and **Supplementary Note 1**). We assigned a phasing score (PQ) to each SNV by comparing the likelihood of the observed data in both phasing states of the variant (Online Methods).

We determined the N50, an indicator of haplotyping performance, from our linked-read analysis. This phasing metric reflects the haplotype block interval at which larger blocks represent 50% of the overall phased genome sequence[1,9]. We identified N50 phase block lengths ranging from 0.9 Mb to 2.8 Mb, with >97% of SNVs phased (**Table 1** and **Fig. 2b**).

Long switch errors occur when a variant position is incorrectly phased in the context of the adjacent flanking variants[9,12]. We used this well-established metric as an indicator of the genome-wide accuracy of our haplotyping analysis (Online Methods). The overall long switch error rate was <0.03% (**Table 1**) for the nuclear trio when compared to phasing information generated via trio sequencing[15]. To further determine phasing accuracy, we examined the distance between all pairwise SNVs for a given haplotype and the probability that these pairs were accurately assigned. Using this metric, phasing accuracy remained >95% out to an interval distance >0.5 Mb (**Fig. 2c**).

We investigated the impact of lower sequencing coverage on phasing performance. We determined the haplotypes of NA12878 using down-sampled proportions of the linked-read data. At approximately 30 Gb of sequence (~10× average coverage), 93% of SNVs were phased, the N50 phase block length was 1.1 Mb and long switch error rates were <0.03% (**Supplementary Table 4**).

Phasing performance decreases in regions with small numbers of heterozygous variants. We found that the probability of correctly phasing SNVs scaled with the pairwise distance between variants (**Fig. 2c**). We examined the PQs at low-density (5 SNVs per 100 kb) heterozygous SNV positions derived from precalled gold-standard variants of NA12878 (ref. 15). We observed only 1,407 variants that fit this criterion, corresponding to <1% of all the heterozygous SNVs. From this set of sparsely distributed SNVs, 94% were confidently phased past the threshold score of PQ > 23. In addition, >60% of the low-density variants were phased with a maximal PQ of 255; this represents exceptionally high phasing quality even for genomic regions with a sparse distribution of SNVs. As a reference, we also examined all heterozygous SNVs in NA12878. We found that 99% were phased accurately and exceeded the threshold score of PQ > 23, and 93% were phased with a maximal, highest-quality PQ of 255. We can potentially increase the phasing performance in low-density heterozygous regions by increasing input molecule length or sequencing depth, or by decreasing the number of genomes loaded per partition.

As another comparison, we analyzed the genome of HapMap subject NA20847, an individual of Gujarati Indian descent. This genome had been sequenced and haplotyped using a fosmid approach[1]. Library quality and phasing performance were similar to results obtained on the nuclear trio (**Table 1**). More than 98% of SNVs were phased, with N50 phase block size of ~2.5 Mb. We calculated a long switch error rate of <0.09% when compared to fosmid-based haplotyping study.

Exome-based phasing

Generating barcode exome libraries is straightforward and enables the haplotyping of genes with linked-reads (Online Methods and **Fig. 1b**). Using ~1 ng of DNA from NA12878, NA12877 and NA12882, barcode libraries underwent exome enrichment with Agilent SureSelect kits and were sequenced to a depth of greater than 185× (**Table 1**). After data preprocessing and alignment, we observed that >99% of the linked reads aligned to the human genome reference, >57% of the bases were on target, and the data had a PCR duplication rate up to 18% at 450× (**Table 1**). Our algorithm phased >95% of genes under 100 kb, and we observed N50 phase block lengths >103 kb for the nuclear trio (**Table 1** and **Supplementary Fig. 2e**,**f**). In addition, haplotype blocks were consistent with Mendelian inheritance across the trio (**Fig. 2d**).

Discovery of phased germline structural variants

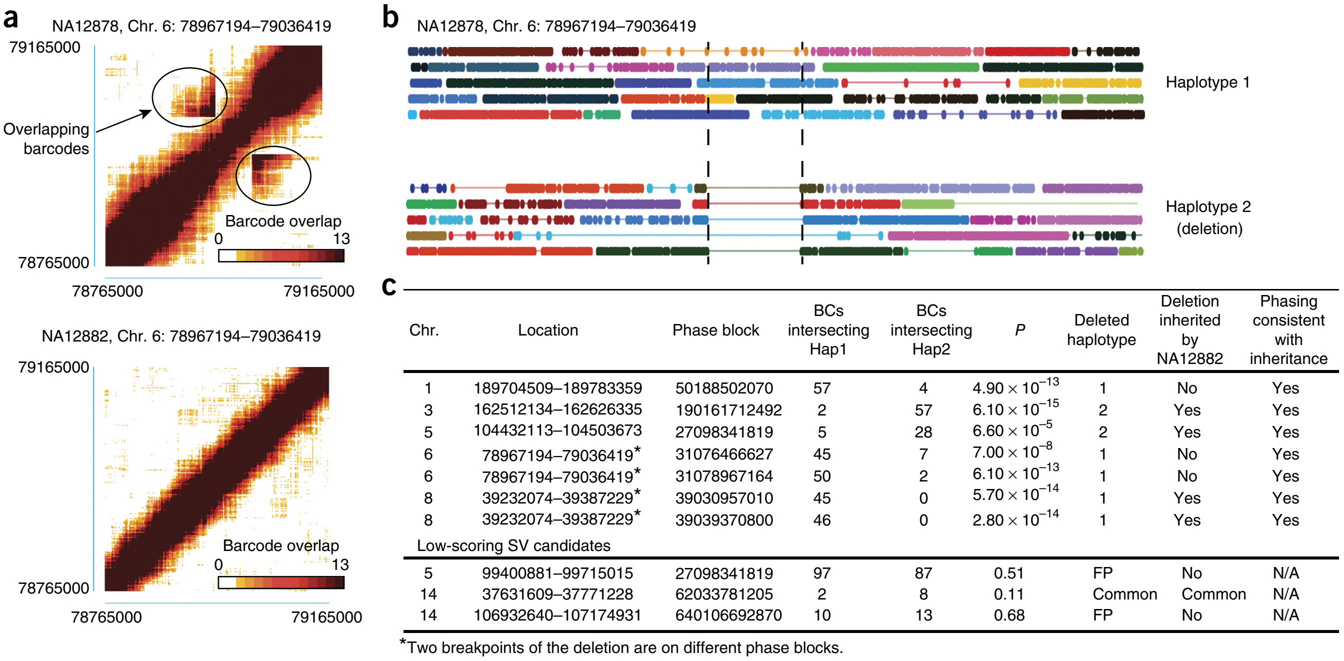
With traditional short-read sequencing approaches, the discovery of structural variants (SVs) is computationally difficult, particularly in highly repetitive regions of the genome. We used linked-read data to call breakpoints in large-scale SVs and assign them to specific haplotypes. We developed an algorithm that efficiently searches all pairs of genomic loci for regions of large numbers of overlapping barcodes. We identified nonadjacent candidate positions of structural variation (**Fig. 3a**,**b** and **Supplementary Fig. 4**). Each breakpoint is assigned a Phred-like quality score, which describes the likelihood of the breakpoint being called by chance (Online Methods and **Supplementary Note 2**). Linked-read data provide sequencing information spanning the region around a breakpoint up to tens of kilobases or more. Thus, linked-read analysis can differentiate between a putative breakpoint created by a true SV and a false positive confounded by repetitive sequences (**Fig. 3b**).

As an initial test, we examined eight large genomic-deletion candidates with sizes greater than 70 kb as previously identified and validated in NA12878 (refs. 16–18). Five were highly ranked by our prediction algorithm, and three had lower scores (**Fig. 3c**). Analysis of barcode count showed that the five high-scoring breakpoints were also consistent with a loss of heterozygosity (LOH), as would be expected for a deletion (**Supplementary Fig. 5**).

We phased these deletions by overlapping the barcodes supporting a breakpoint with barcodes from neighboring haplotype blocks (**Fig. 3b**). Counting the haplotype-specific barcodes provides another type of score for vetting of the putative deletion. Five of these candidate deletions had flanking haplotype blocks that cover both sides of the deletion (**Fig. 3c**). In contrast, the breakpoints of two putative candidates with a low score could not be phased. As additional evidence of the accuracy of our haplotype deletion calls, we examined the consistency of haplotype blocks with Mendelian inheritance for NA12878 (the mother) and NA12882 (the child). Among the five deletions with the highest scores, three were inherited in the child (**Fig. 3c** and **Supplementary Table 5a**,**b**). For the highest ranking cases, when the analysis of the child showed the deletion, it also showed the related haplotype; when the child did not inherit the deletion, the other maternal haplotype was inherited (**Fig. 3c** and **Supplementary Table 6**).

We validated these breakpoints with targeted sequencing[19,20]. We used targeting probes that tile across the genomic intervals where the two breakpoints are created by a deletion (Online Methods, **Supplementary Table 7** and **Supplementary Fig. 6**). We examined targeted reads for evidence of a breakpoint junction and deletion. In general, the chimeric junctions of the high-scoring deletion breakpoints were validated and showed the appropriate inheritance pattern (**Supplementary Fig. 6** and **Supplementary Table 6**). In contrast, two of the lower-scoring deletion candidates could not be confirmed, one of which was in a V(D)J recombination region (**Supplementary Table 6**). The other lower-scoring candidates had targeted sequencing results pointing to a deletion that was seen among all three individuals.

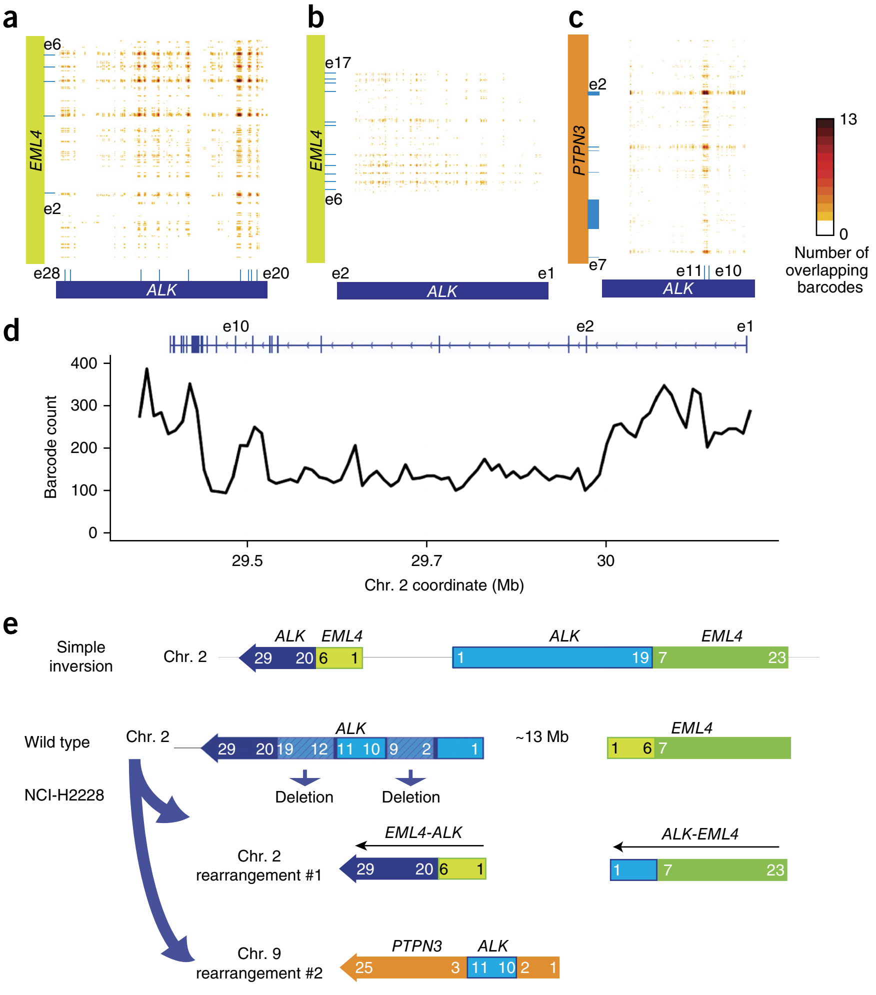
Linked-read data reveal other types of structural rearrangements besides deletions. Overall, our SV algorithm called 20 structural variants in NA12878, of which 11 were identified as deletions in a recent *de novo* assembly[10], two were reported as inversions[16] and one was identified as a retrotransposon insertion[21] (**Supplementary**



Chr.	Location	Phase block	BCs intersecting Hap1	BCs intersecting Hap2	*P*	Deleted haplotype	Deletion inherited by NA12882	Phasing consistent with inheritance
1	189704509–189783359	50188502070	57	4	4.90×10^{-13}	1	No	Yes
3	162512134–162626335	190161712492	2	57	6.10×10^{-15}	2	Yes	Yes
5	104432113–104503673	27098341819	5	28	6.60×10^{-5}	2	Yes	Yes
6	78967194–79036419*	31076466627	45	7	7.00×10^{-8}	1	No	Yes
6	78967194–79036419*	31078967164	50	2	6.10×10^{-13}	1	No	Yes
8	39232074–39387229*	39030957010	45	0	5.70×10^{-14}	1	Yes	Yes
8	39232074–39387229*	39039370800	46	0	2.80×10^{-14}	1	Yes	Yes
Low-scoring SV candidates								
5	99400881–99715015	27098341819	97	87	0.51	FP	No	N/A
14	37631609–37771228	62033781205	2	8	0.11	Common	Common	N/A
14	106932640–107174931	640106692870	10	13	0.68	FP	No	N/A

*Two breakpoints of the deletion are on different phase blocks.

Figure 3 Detecting genomic deletions in NA12878. (**a**) Top, heat map of overlapping barcodes plotted for a deletion on chromosome 6 (chr. 6): 78967194–79036419 in NA12878. Black circle indicate overlapping barcodes near the breakpoints. Bottom, heat map of barcodes in the same region for NA12882, shown as a negative control. (**b**) Linked-read data of NA12878 WGS sample spanning chr. 6: 78967194–79036419. In haplotype 1 (top), overlapping barcodes are observed only in contiguous regions. In haplotype 2 (bottom), a deletion is shown as a gap in linked reads. In contrast to regions without a deletion, barcodes in the region before the gap overlap with barcodes in the region after the gap. Horizontal lines represent linked reads with the same barcode; dots represent reads; colors indicate barcodes. Dashed vertical black lines represent breakpoints. (**c**) Summary of eight deletion candidates, including supporting evidence from overlapping barcode (BC) count, phasing of the deletion breakpoints and inheritance support in NA12882. Whereas the five high-scoring SV candidates have support from each type of evidence, two lower-scoring candidates lack support from any evidence including targeted sequencing. Haplotype assignment in one phase block is not necessarily the same as the haplotype assignment in a different phase block. Hap1, haplotype 1; hap2, haplotype 2.



Gene fusion	Breakpoint 1	Breakpoint 2	Phase block location	BCs overlapping breakpoint	BCs intersecting haplotype 1	BCs intersecting haplotype 2	P	Haplotype involved in gene fusion
ALK-EML4	Chr. 2: 29435765	Chr. 2: 42471639	Chr. 2: 29225502–29446202	25	23	0	2.4×10^{-7}	1
ALK-PTPN3	Chr. 2: 29489574	Chr. 9: 112225531	Chr. 9: 112219474–112225774	30	0	21	9.5×10^{-7}	2

Figure 4 Rearrangement detection of an *EML4-ALK* fusion from exome sequencing of NCI-H2228. (**a**) Overlap of barcodes between exons 20–28 (e20, e28) of *ALK* and exons 2–6 (e2–e6) of *EML4*. (**b**) Overlap of barcodes between e1–e2 of *ALK* and e7–e17 of *EML4*. (**c**) Overlap of barcodes between e10–e11 of *ALK* and the 5′ end of *PTPN3*. Blue bars in **a**–**c** represent exons. (**d**) Barcode counts in *ALK* of NCI-H2228 WGS sample. (**e**) Schematics of complex chromosomal rearrangement involving *ALK*, *EML4* and *PTPN3*. Instead of the simple inversion reported in the literature, we observed a deletion, an inversion of *ALK* on chromosome 2 with *EML4* and an insertion of *ALK* into *PTPN3* on chromosome 9. (**f**) Phasing support around *ALK* and *PTPN3* breakpoints in *EML4-ALK* and *ALK-PTPN3* gene fusion. Haplotype assignment in one phase block is not necessarily the same as the haplotype assignment in a different phase block. Chr., chromosome.

Fig. 4, **Supplementary Table 8** and Online Methods). To assess the sensitivity of our method, we compared our calls to a set of deletions identified from both conventional short-read sequencing[18] and long-read–assisted assembly[10]. There were only three such deletions, all of which were detected and validated (**Fig. 3c** and Online Methods).

Detection of an *EML4-ALK* rearrangement via exome phasing

SVs such as cancer rearrangements frequently occur in intronic sequences rather than exons and can lead to chimeric gene products. Exome sequencing does not detect gene fusions for which the breakpoint is more than a few hundred base pairs from an exon without custom targeting assays and extremely high sequencing coverage[22,23]. To overcome these issues, we used exome linked reads to detect a clinically actionable cancer rearrangement. The lung cancer cell line NCI-H2228 contains an *EML4-ALK* fusion[24,25], in which exons 1–6 of *EML4* are fused to exons 20–29 of *ALK*. This rearrangement leads to constitutive activation of ALK[26], an oncogenic kinase driver. We prepared a barcode sequencing library from ~1 ng of genomic DNA, conducted exome enrichment and sequenced to an average sequencing coverage of 204× after duplication filtering (**Table 1**).

We correctly identified an *EML4-ALK* fusion (**Fig. 4a**–**d**, **Supplementary Fig. 7a**,**b** and **Supplementary Table 9**); our exome linked-read data showed that the rearrangement occurs between exons 20–26 of *ALK* and exons 2–6 of *EML4* (**Fig. 4a**), consistent with previous reports and our own validation (**Supplementary Fig. 7**). A simple inversion would predict corresponding overlap between exon 19 of *ALK* with exon 7 of *EML4* (**Fig. 4e**). Our results showed overlap of exon 1 of *ALK* and exon 7 of *EML4* (**Fig. 4b**), suggesting a deletion of exons 2–19 of *ALK* and a more complex structure than a simple inversion. In addition, we identified an additional insertion of *ALK* exons 10–11 in the gene *PTPN3* on chromosome 9 (**Fig. 4c**, **Supplementary Fig. 7c**,**d** and **Supplementary Table 9**) as has been previously reported[27].

On the basis of these results for this cell line, we inferred a refined structure of the overall structural rearrangement (**Fig. 4e**) covering the *ALK* deletion, *EML4-ALK* inversion and insertion of exons 10–11 of *ALK* into *PTPN3*. Exons 20–29 of *ALK* are contained in a 220-kb phase block; only one haplotype overlaps with the *EML4-ALK* fusion. Similarly, exons 3–4 of *PTPN3* are contained with a 40-kb phase block, and there is a distinct segregation of the *ALK* insertion into only one haplotype of the *PTPN3* gene (**Fig. 4f**). The rearrangement structure was separately verified with linked-reads WGS (**Supplementary Table 1** and **Supplementary Fig. 7c**,**d**). Analysis of the barcode counts in the WGS data (**Fig. 4d**,**f**) revealed a coverage reduction consistent with a deletion in the region covering exons 2–19 of *ALK*.

Haplotype analysis of a primary colon adenocarcinoma

Whole-genome haplotype analysis has been reported for the HeLa tumor cell line[2], but phasing analysis of primary tumor genomes has remained difficult. Unlike cancer cell lines, which are more homogeneous in their genetic composition, primary tumors derived from clinical samples pose a number of challenges. The amount of DNA from clinical samples is often limited, and the cellular composition

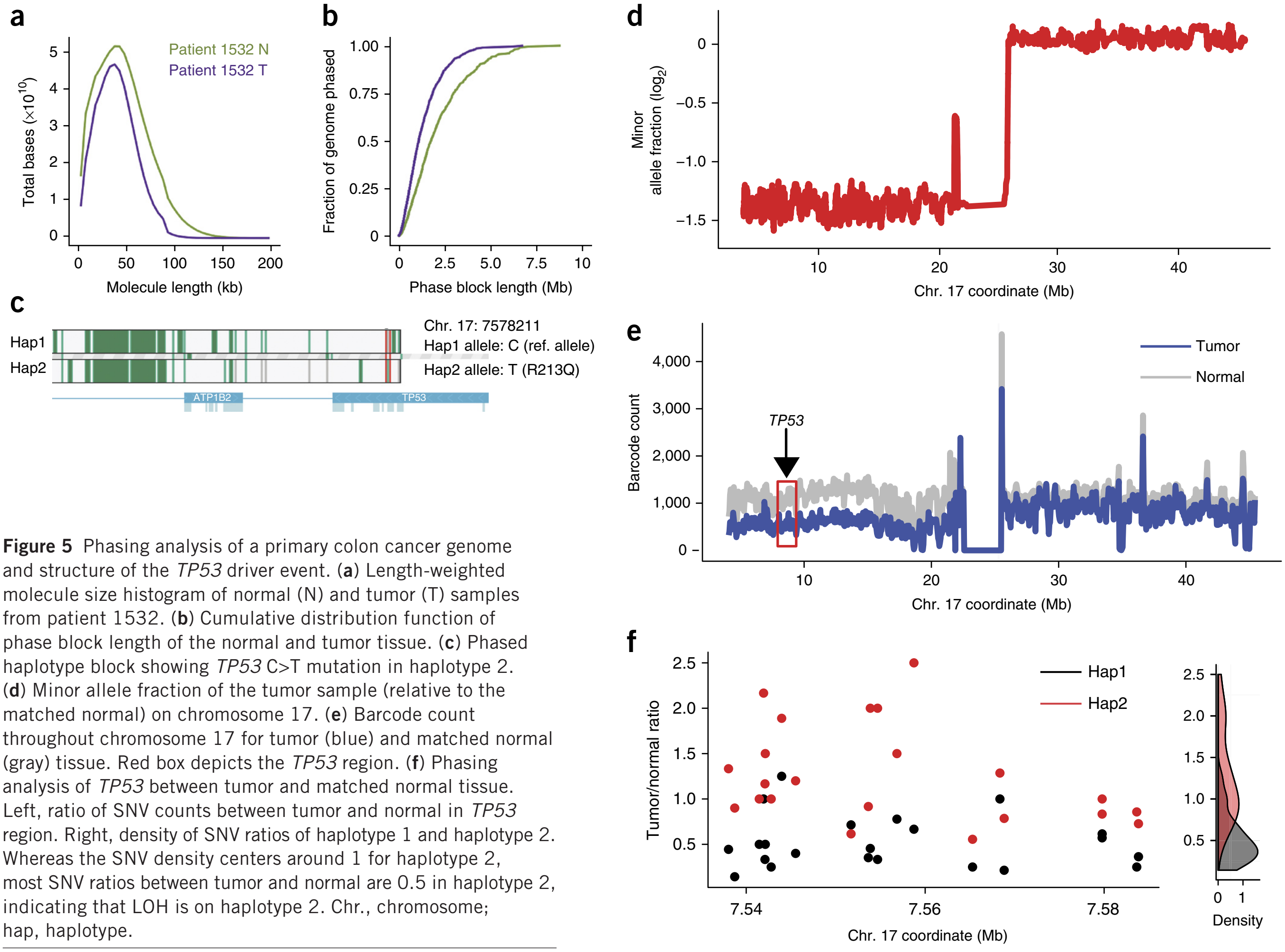


Figure 5 Phasing analysis of a primary colon cancer genome and structure of the *TP53* driver event. (**a**) Length-weighted molecule size histogram of normal (N) and tumor (T) samples from patient 1532. (**b**) Cumulative distribution function of phase block length of the normal and tumor tissue. (**c**) Phased haplotype block showing *TP53* C>T mutation in haplotype 2. (**d**) Minor allele fraction of the tumor sample (relative to the matched normal) on chromosome 17. (**e**) Barcode count throughout chromosome 17 for tumor (blue) and matched normal (gray) tissue. Red box depicts the *TP53* region. (**f**) Phasing analysis of *TP53* between tumor and matched normal tissue. Left, ratio of SNV counts between tumor and normal in *TP53* region. Right, density of SNV ratios of haplotype 1 and haplotype 2. Whereas the SNV density centers around 1 for haplotype 2, most SNV ratios between tumor and normal are 0.5 in haplotype 2, indicating that LOH is on haplotype 2. Chr., chromosome; hap, haplotype.

frequently includes normal stromal tissue. As a first-time demonstration of a phasing analysis for a primary cancer genome, we analyzed a primary colon adenocarcinoma and identified three classes of genetic aberration in the context of megabase-scale haplotypes: (i) mutations, (ii) copy number variants (CNVs) and (iii) rearrangements. First, we performed short-read WGS on both the tumor and matched normal pair to an average sequencing coverage of 50× followed by variant calling (**Supplementary Table 1**). For both samples, we used ~2 million heterozygous SNVs determined from the short reads for subsequent phasing (**Supplementary Table 3** and Online Methods)**.**

For the phased analysis of a primary cancer genome, we used ~1 ng genomic DNA from the tumor and normal sample as direct inputs, without other preparative steps, for generating a barcode sequencing library. This tumor had 70% purity. We sequenced the barcode libraries to an average coverage of ~30×. After linked-read alignment, we determined that the genomic DNA had an inferred DNA length distribution with mean >50 kb (**Fig. 5a**). More than 94% of SNVs were phased, with N50 phase block lengths of ~1.5 Mb for the normal sample and ~0.9 Mb for the tumor (**Table 1** and **Fig. 5b**). We generated haplotype blocks for both samples and examined the haplotype intersection intervals. From this intersection, 90% of the bases were shared between the two samples.

Seventeen deleterious cancer mutations were identified per Combined Annotation Dependent Depletion (CADD) scores[28] and assigned to specific haplotype blocks (**Supplementary Table 10**). A number of the mutations occurred in known colorectal cancer drivers such as *TP53* and *NRAS*[29]. Using linked-read analysis, we identified five rearrangements. Two were interchromosomal translocations (**Supplementary Table 11**). The short-read WGS data provided validation of these events; the breakpoints of the five structural variants were confirmed by BreakDancer[30] predictions and supported by reads covering the breakpoints (**Supplementary Table 11**). Analysis of the short-read WGS for copy-number alteration[31] identified 26 CNV intervals, of which 24 were validated by counting the average number of barcodes and linked reads (**Supplementary Table 12**). Two CNV intervals (short-read analysis) were not validated by barcode counting and occurred in centromeric and telomeric regions that had lower linked-read coverage.

We used linked-read analysis and haplotype blocks to explore the context and structural alterations affecting a critical driver mutation in *TP53*, namely a C>T mutation encoding a candidate deleterious nonsynonymous R213Q substitution (**Supplementary Table 10**). Phasing analysis demonstrated that the C>T mutation is within a 46-kb phase block on the haplotype 2 allele (**Fig. 5c**). Traditional short-read WGS analysis indicated an LOH event represented by a hemizygous deletion in chromosome 17p (**Fig. 5d**). Barcode count analysis with linked reads confirmed this observation; given the mutation allelic fraction, the corrected copy number at that region is 1. The LOH results from an extensive genomic deletion overlapping the *TP53* mutation (**Fig. 5e**). The phased SNV frequencies in the haplotype 1 allele were lower in the tumor than in the normal sample, indicating that LOH in the tumor sample is associated with the loss of the haplotype 1 allele (**Fig. 5f**). Thus, this *TP53* C>T mutation is in *trans* with the deleted allele haplotype. As a result, the tumor contains only a single, inactivated copy of *TP53*. Taken together, this result shows the unambiguous biallelic inactivation of *TP53* (refs. 32,33).

DISCUSSION

We demonstrate the use of a high-throughput microfluidics technology to construct phased sequencing libraries from nanogram inputs of HMW DNA. By using gel beads as the barcode delivery reagent, we demonstrate the robust loading of a single barcode to a microdroplet partition. This technology addresses issues that affect current experimental phasing approaches, such as problems of Poisson-distributed barcode loading and limited partitioning with microwells. To our knowledge, this is the first study to demonstrate a droplet-based system for whole-genome phasing and structural variant analysis. In addition to phasing and structural variant calling, linked reads can potentially also be applied to *de novo* genome assembly, remapping of difficult regions of the genome, detection of rare alleles and elucidation of complex structural rearrangements.

Several studies have demonstrated high-throughput barcoding of droplet partitions[34–36] for single-cell RNA-seq and analysis of short bacterial 16S sequences. These approaches use up to millions of unique barcodes that are then incorporated into partitions. However, none of these droplet applications generate megabase-scale haplotypes from WGS. There are a number of other genome sequencing approaches used for phasing[1,2,5–9,37,38] (an overview is listed in **Supplementary Table 13**). Only one of these approaches uses droplets; this method does not involve sequencing but rather relies on digital PCR counting methods to assess a singleplex candidate locus[38].

To assess performance, we conducted a phased genome analysis on several well-defined genomes. With this technology, we phased more than 95% of SNVs in all samples with N50 phase block sizes ranging from 0.8 Mb to 2.8 Mb, at a low switch error rate (<0.001). This phasing performance was achieved with existing variant data sets. We show that linked-read data can be used to phase *de novo* variants, although more coverage will be required to achieve parity with standard library preparation methods owing to coverage biases against GC-rich regions (**Supplementary Fig. 2d**). Statistical inference of haplotypes from genomic intervals dominated by similar heterozygous variants among family members is an issue that experimental phasing overcomes. For example, in the NA12878 nuclear trio, about 10% of the total number of SNVs in the child were inherited from such regions with common genotypes[39].

Our technology is compatible with standard downstream NGS assays such as exome enrichment, as barcode information is introduced as the first step in the library preparation process. With the nuclear trio samples, more than 95% of genes <100 kb long were phased with this phased exome sequencing approach, which enables the economical use of phased analysis on large numbers of samples.

We used phasing and read barcode counts to identify structural variations such as large genomic deletions and rearrangements, which were independently validated by multiple methods. Using exome linked reads, we delineated the complex rearrangements such as the *EML4-ALK* inversion in the NCI-H2228 cell line. In addition, we showed that linked-read phasing of structural variants distinguishes true SVs from false predictions.

We used this approach to phase a cancer genome derived from a primary tumor. The combination of somatic mutations, haplotype blocks and barcode counting identified the *trans*- relationship between a mutation in *TP53* and a chromosome 17p loss in colon adenocarcinoma. We also generated haplotypes incorporating other critical genetic aberrations, such as copy number alterations and rearrangements. We anticipate that phased cancer genomes will provide new insight into the genomic alterations underlying tumor development and maintenance.

The identification of potentially pathogenic mutations and structural variants remains a challenge, and linked-read sequencing provides an opportunity to improve the understanding of diseases such as cancer.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Data have been deposited in the Short Read Archive (SRA) under accession number SRP051629 and in dbGAP under accession number phs000898.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health grants NHGRI P01HG000205 (to B.T.L., E.S.H., S.M.G., J.M.B. and H.P.J.), NCI R33CA174575 (to J.M.B., S. Greer and H.P.J.) and NHGRI R01HG006137 (to H.P.J.). The American Cancer Society provided additional support to S. Greer and H.P.J. (Research Scholar grant, RSG-13-297-01-TBG). H.P.J. also received support from the Doris Duke Clinical Foundation, the Clayville Foundation, the Seiler Foundation and the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

B.T.L., M.S.-L., M.J., J.M.B., C.M.H., S.K.-P., L. Merrill, R.B., A.J.M., Y.L., A.D.P., A.J.L., P.H., L.G., K.B., P.P., E.S.H., C.W., K.M.G., S.S., K.D.N., B.J.H. and H.P.J. designed the experiments. B.T.L., J.M.B., C.M.H., L. Merrill, J.M.T., P.A.M., P.W.W., R.B., A.J.M., Y.L., P.B., A.D.P., A.J.L., P.M., G.M.V., L. Montesclaros, M.L., L.G., D.E.B., K.B., P.P., E.S.H., C.W., J.P.D., I.W., H.S.O, J.Y.L., Z.W.B., K.M.G, G.P.M., Z.W.B., F.M., N.O.K., J.A.B., S.G., C.B., A.N.F., A.C. and B.J.H. conducted the experiments. D.A.M., R.B., A.J.M., S.W.S., S.K., J.A.B., A.K., K.D.N. and B.J.H. designed the instrument. M.S.-L., M.J., C.M.H., P.W.W., R.B., A.J.M., Y.L., A.D.P., A.J.L., P.H., L. Merrill, L.G., K.P.B., P.P., S.K., J.P.D., J.A.B., K.D.N. and B.J.H. designed reagents for phasing. B.T.L, J.M.B., E.S.H. and H.P.J. designed reagents for targeted sequencing analysis. G.X.Y.Z., M.S.-L., S.K.-P., P.M., G.K.L., D.L.S., W.H.H., R.T.W., S.S. and K.D.N. wrote the haplotype analysis algorithms. J.M.B. and S.M.G. wrote the analysis algorithms for short-read sequencing analysis. M.S.-L., P.J.M, A.W., G.K.L., D.L.S., W.H.H. and R.T.W. wrote the analysis software. G.X.Y.Z., B.T.L., M.S.-L., M.J., J.M.B., C.M.H., S.K.P., J.M.T., R.B., A.J.M., Y.L., P.B., P.M., P.H., L. Merrill, M.L., A.W., K.B., P.P., S.K., J.P.D., I.W., H.S.O., S.M.G., S. Greer, J.Y.L., Z.W.B., K.M.G., W.H.H., G.P.M., Z.W.B., F.M., J.A.B., S. Gauby, C.B., A.N.F., W.H.H., A.C., S.S., K.D.N., B.J.H. and H.P.J. analyzed the data. G.X.Y.Z., B.T.L., M.S.-L., M.J., S. Greer, B.J.H. and H.P.J. wrote the manuscript. H.P.J. oversaw the overall genetic experiments and analysis.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
2. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).
3. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
4. Suk, E.K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
5. Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res.* **40**, 2041–2053 (2012).

6. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
7. Kaper, F. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 5552–5557 (2013).
8. Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
9. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
10. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
11. Abate, A.R., Chen, C.H., Agresti, J.J. & Weitz, D.A. Beating Poisson encapsulation statistics using close-packed ordering. *Lab Chip* **9**, 2628–2631 (2009).
12. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* **32**, 261–266 (2014).
13. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
15. Cleary, J.G. *et al.* Joint variant and *de novo* mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.* **21**, 405–419 (2014).
16. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
17. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
18. Mills, R.E. *et al.* 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
19. Hopmans, E.S. *et al.* A programmable method for massively parallel targeted sequencing. *Nucleic Acids Res.* **42**, e88 (2014).
20. Myllykangas, S., Buenrostro, J.D., Natsoulis, G., Bell, J.M. & Ji, H.P. Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing. *Nat. Biotechnol.* **29**, 1024–1027 (2011).
21. Schrider, D.R. *et al.* Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* **9**, e1003242 (2013).
22. Frampton, G.M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
23. Lipson, D. *et al.* Identification of new *ALK* and *RET* gene fusions from colorectal and lung cancer biopsies. *Nat. Med.* **18**, 382–384 (2012).
24. Choi, Y.L. *et al.* Identification of novel isoforms of the *EML4-ALK* transforming gene in non-small cell lung cancer. *Cancer Res.* **68**, 4971–4976 (2008).
25. Koivunen, J.P. *et al.* *EML4-ALK* fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin. Cancer Res.* **14**, 4275–4283 (2008).
26. Soda, M. *et al.* Identification of the transforming *EML4-ALK* fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).
27. Jung, Y. *et al.* Discovery of *ALK-PTPN3* gene fusion from human non-small cell lung carcinoma cell line using next-generation RNA sequencing. *Genes Chromosom. Cancer* **51**, 590–597 (2012).
28. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
29. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
30. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
31. Shen, J.J. & Zhang, N.R. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.* **6**, 476–496 (2012).
32. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
33. Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.* **319**, 525–532 (1988).
34. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
35. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
36. Borgström, E. *et al.* Phasing of single DNA molecules by massively parallel barcoding. *Nat. Commun.* **6**, 7173 (2015).
37. de Vree, P.J. *et al.* Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.* **32**, 1019–1025 (2014).
38. Regan, J.F. *et al.* A rapid molecular approach for chromosomal phasing. *PLoS ONE* **10**, e0118270 (2015).
39. Roach, J.C. *et al.* Chromosomal haplotypes by genetic phasing of human families. *Am. J. Hum. Genet.* **89**, 382–397 (2011).

ONLINE METHODS

Genomic DNA samples. The Institutional Review Board (IRB) at Stanford University School of Medicine approved this study. Informed consent was obtained, and the samples were made available from the Stanford Cancer Center Tissue Bank. This study used a primary colorectal adenocarcinoma and matched normal tissue that were collected and flash frozen at time of surgical resection. Both samples had genomic DNA extracted with the E.Z.N.A. SQ DNA/RNA Protein Kit (Omega Bio-Tek). The genomic DNA did not require further size selection or processing. We quantified the DNA with Life Technologies Qubit.

For the commercially acquired genomic DNA, we size selected DNA molecules 20 kb or larger using the BluePippin (Sage Science) (NA12877 and NA12882 from Coriell, and NCI-H2228 from ATCC). In addition, we harvested immortalized human lymphocyte cells (GM12878 and GM20847 from Coriell), and genomic DNA was extracted using the Gentra Puregene Cell kit (Qiagen). All cell lines have been tested for mycoplasma contamination.

Sequencing library construction using the GemCode platform. A GemCode Instrument (10X Genomics) was used for sample preparation. The high-throughput nature of the platform allows construction of eight sequencing libraries by one person in a day. Sample indexing and partition barcoded libraries were prepared using a beta version of the GemCode Gel Bead and Library Kit (10X Genomics, Pleasanton, CA). 1 ng of sample DNA was used for GEM reactions where DNA molecules were partitioned into droplets to amplify the DNA and introduce 14-bp partition barcodes. With 1 ng genomic DNA of 50 kb molecule length, there are ~100 molecules per droplet.

GEM reactions were thermal cycled (95 °C for 5 min; cycled 18×: 4 °C for 30 s, 45 °C for 1 s, 70 °C for 20 s, and 98 °C for 30 s; held at 4 °C). After amplification, the droplets were fractured and the library intermediate DNA purified using the 10X Genomics protocol. The DNA was subsequently sheared to either 250 bp or 500 bp using a Covaris M220 system (**Supplementary Table 2**) to construct sample-indexed libraries using 10X Genomics adapters. The barcode sequencing libraries were quantified by quantitative PCR (qPCR) (KAPA Biosystems Library Quantification Kit for Illumina platforms). Sequencing was conducted with an Illumina Hiseq2500 with 2 × 98 paired-end reads based on the manufacturer's protocols.

To compare barcode libraries against standard short-read libraries, we prepared a TruSeq library (Illumina) following manufacturer's protocols, using 100 ng of DNA. Both barcode and TruSeq libraries used GM12878 genomic DNA. Each library was sequenced to ~30× coverage. At 30× coverage, the coverage of molecule in each droplet is 0.1×, and the number of linked reads per molecule is around 15.

Five micrograms of each barcode library was used for exome capture (Agilent SureSelect Human All Exon V5+UTRs) with the Agilent SureSelect Target Enrichment System (Agilent Technologies, Santa Clara, CA) supplemented with modified blocking oligonucleotides for Illumina Dual Indexing (TS HT i5 and TS HT i7) from IDT. Captured libraries were quantified by qPCR (KAPA Biosystems). Again, sequencing was conducted with an Illumina Hiseq2500 with 2 × 98 paired-end reads based on the manufacturer's protocols.

Alignment, barcode assignment and calculation of sequencing metrics. The GemCode analysis software was used for processing the sequenced data from barcode libraries. Fastq files from Illumina sequencing reads were trimmed (removing the first 10 nt of all reads) and aligned to the human genome (hg19) using bwa (mem algorithm, version 0.7.10-r789). Barcodes were incorporated into the read information in the BAM file, and only reads associated with valid barcodes were considered for alignment and downstream analysis. For visualization and some analysis, barcode counts were calculated using nonoverlapping window size of 100 kb, over all positions. Only uniquely mapped, nonduplicated reads with mapping quality (MAPQ) of 60 were considered.

Reads were sorted by position using Samtools (version 0.1.19-96b5f2294a). PCR duplicates were marked if two sets of read pairs shared both identical aligned genomic position and an identical associated barcode sequence. Linked reads were inferred by clustering reads from the same barcode on the genome, and their boundaries were set by two nearest reads more than 50 kb apart. The term 'barcodes correctly assigned' is the fraction of barcodes matching a known barcode. Relative genomic loading per partition was calculated as the fraction of the amount of DNA in a partition relative to the size of the human genome. The number of binding events is estimated as the product of binding density and genome loaded per partition.

For a uniform distribution of barcode frequencies, the probability of drawing two identical barcodes is $p = \sum_i frequency(BC_i) * frequency(BC_i) = N/N^2 = 1/N$ where N is the number of unique barcodes. Thus, effective barcode diversity, which accounts for a nonuniform distribution of barcode frequencies, can be calculated as:

$$Effective\, barcode\, diversity = \frac{1}{\sum_i frequency(BC_i) * frequency(BC_i)}$$

where BC_i = i-th barcode.

To perform the variant calling analysis, we used Freebayes to call variants on 10X and TruSeq libraries, down-sampling each library to 10×, 20× and 30× coverage. Then sensitivity and PPV of SNVs was evaluated against ground-truth variants published by Cleary *et al.*[15].

Phasing linked reads. See **Supplementary Note 1**.

Structural variant calling from linked-read data. See **Supplementary Note 2**.

Phasing of structural variants. Phasing of large-scale variants used the final probabilistic assignment of barcodes to haplotype blocks calculated as part of the phasing code. For each haplotype block within a 30 kb window of each of the two breakpoints defining a structural variant, barcodes supporting the structural variant call were assigned to one of the two haplotypes for that haplotype block. For each haplotype block, the counts of barcodes assigned to each of the two haplotypes were used to calculate a *P* value under the two-tailed binomial test. Phase calls were made on a structural variant when the $P < 0.01$.

Validation of genomic deletions with targeted sequencing. We validated a series of genomic deletions using targeted sequencing[20]. The methods are fully described by Hopmans *et al.*[19]. For this validation study, we relied on targeting assays that use target-specific primer probes that hybridize to the target DNA molecule[20]. Afterward, a polymerase extension captures the specific genomic target sequence. Previously, we demonstrated the utility of this method for confirming SVs, even in the context of genomic mixtures where a candidate rearrangement is present in only a fraction of the sample[19]. As a result of random fragmentation of genomic DNA in the library preparation, breakpoints of structural variants will be randomly distributed within a subset of the sequencing reads.

For this assay, we designed multiple primer-probe sequences flanking each putative breakpoint associated with a structural variant candidate. This targeting method is generally successful at selecting sequences up to 1 kb or farther from the primer probe. The primer-probe sequences chosen were on both the forward and reverse strands surrounding both sides of a target putative breakpoint within a distance of 0.75 kb (**Supplementary Fig. 6**). Reads captured by primer probes upstream from the breakpoints should cross on the reverse strand; reads captured by primer probes downstream from the breakpoints should cross on the forward strand.

For the eight candidate deletions that were validated, we designed and synthesized 163 primer-probe oligonucleotides (**Supplementary Table 7**). Generally, all of these oligonucleotides were unique in terms of their representation in the genome. The only exception was for 15 probes intended to validate a deletion in chromosome 5 (position 99400335–99713992). This deletion occurs in an area of the genome that is highly repetitive, so only two of these 15 primer probes contain a 20-mer that aligns uniquely to the human genome with no single-mismatch alignments.

Single-end alignment using bwa (mem algorithm, version 0.7.10) was performed on the individual reads from the mate pairs. The targeting primer sequence is included in read 2 and used as an index for a given target segment. The captured sequence is in read 1 and were indexed on the basis of the read 2 targeting primer. The read 1 sequences that completely aligned to the human genome were excluded. The remaining read 1 sequences were evaluated for evidence of breakpoint and counted. The reads that had breakpoints were concatenated to create a breakpoint sequence. Reads crossing breakpoints were generated by finding reads that included a soft-clipped section such that the aligning portion preceded or followed the breakpoint; soft-clipped reads that also included soft-clipping on the non-breakpoint side were excluded. Using this read set we counted 20-mers that contained a chimeric junction containing sequence on both sides of the breakpoint candidate.

Evaluation of SV calls in NA12878. To assess the false discovery rate of our SV calling algorithm, we compared our SV calls in NA12878 against a recent *de novo* assembly using genomic DNA from this individual[10]. We obtained a list of assembly-based deletion and insertion calls in NA12878 from the Genome in a Bottle website (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai). We then constructed two deletion data sets: (i) a 'confident' set containing deletions that were marked as 'passing' by the study. These were deletions that were called by three or more out of the seven methods used in that paper[10]; (ii) a 'relaxed' set containing all deletions detected by at least one computational method in the *de novo* assembly data.

We focused on deletion calls in the following comparison because (i) deletion calls are much easier to compare across data sets; (ii) we omitted insertions from the above two sets because our algorithm is not designed to detect gaps in the reference genome. Out of the 20 calls that were made in NA12878 via linked reads, 40% and 55% respectively matched those from the confident and relaxed Pendleton data sets[10] to within 20 kb.

Besides deletion calls, our SV algorithm can detect other types of structural rearrangement. Indeed, two of our calls matched inversions reported in the literature[16]. One additional call is a retrotransposon insertion that has been found in individuals of European descent[21]. Although the three calls were not explicitly called by Pendleton *et al.*[10], they were supported by long sequence reads (i.e. Pacific Biosciences sequencer) obtained from the same assembly work[10]. Altogether, this increases the percentage of the validated calls against the *de novo* assembly to 70%. We have included a comparison of the calls in **Supplementary Table 8**.

RT-PCR validation of *EML4-ALK* fusion. Reverse transcription PCR (RT-PCR) was used to confirm the *EML4-ALK* and *ALK-PTPN3* fusions in NCI-H2228 cancer cell line. We used the Cells-to-CT 1-Step Power SYBR Green Kit (Life Technologies) according to the manufacturer's recommendations. *ACTB* was assayed using SYBR Green Kit Control Kit (Life Technologies). As a negative control, NA12878 cells were assayed in parallel. Briefly, ~7,500 cells were lysed and treated with DNase I in a total of 55 µl. 2 µl of lysate was used for a 20 µl PCR reaction. The PCR products were visualized using the BioAnalyzer High Sensitivity DNA Kit (Agilent), with the amplicons *ACTB* diluted 1:50, *EML4-ALK* 1:20, and *ALK-PTPN3* diluted 1:3. The primers for the *EML4-ALK* amplicon are (F) 5′-GCATAAAGATG TCATCATCAACCAAG; (R) 5′-CGGAGCTTGCTCAGCTTGTA. The PCR primers for *ALK-PTPN3* are: (F) 5′-TGGCTGCAGATGGTCGCATGG; (R) 5′-AGTCCACGGAGTCGTCATCAT.

Cancer WGS with short reads and data processing. Whole-genome libraries were made per the manufacturer's protocol (Illumina). Sequencing libraries underwent cluster generation on an Illumina cBot using paired-end flow cells and Illumina TruSeq chemistry and sequenced at Illumina with the HiSeq 2500 for 2 × 100 cycle reads with indexing. Sequence reads were aligned to the human genome version hg19 using bwa[13]. The Genome Analysis Toolkit (GATK)[14] was used to determine overall sequencing coverage and variant calls.

Cancer genome somatic mutation calling for coding mutations. The WGS data were aligned using bwa 0.7.5 (ref. 13) 'aln' and 'sampe' with default parameters against NCBI human genome build 37. Data were sorted and duplicate marked using Picard's AddOrReplaceReadGroups and MarkDuplicates functions, respectively. Picard version 1.63 was used in all steps. The files were merged in the GATK[14] RealignerTargetCreator step. This step and the IndelRealigner step were used to realign locally; IndelRealigner referred to dbSNP version 135. The BaseRecalibrator function used CycleCovariate and ContextCovariate as covariates and referred to dbSNP 135. At this point the realigned BAM file of patient 1532's data was split up to allow easier processing. GATK PrintReads was run on realigned BAM files with the appropriate recalibration data table to produce recalibrated BAM files. The GATK UnifiedGenotyper was then run with the parameters–dbsnp dbsnp_135.b37.vcf–max_alternate_alleles 11. These raw calls were then recombined. The GATK VariantRecalibrator was run on the raw VCF data, using the hapmap, omni, and dbsnp resources with standard priors and using HaplotypeScore, MQRankSum, ReadPosRankSum, FisherStrand (FS), RMSMappingQuality (MQ) and Coverage (DP) as filter elements. Finally, the ApplyRecalibration step was used to determine whether calls received a PASS value or not. Variants were called using GATK version 2.6-4. After variants were called, all SNV positions where the tumor and normal calls differed were submitted to CADD annotation[28]. SNVs were then filtered to require a somatic variant (positions where the normal tissue shows no variant and the tumor does, or the normal tissue is heterozygous and the tumor has a homozygous variant) in a coding region with coverage depth ≥10 in both samples and a CADD Phred score ≥25 (**Supplementary Table 10**). Sequencing coverage was assessed with the GATK DepthOfCoverage tool at depths of 10, 20 and 30 (**Supplementary Table 1**).

SNVs were then extracted from the phased VCF files, and their phasing status was assessed. The tumor haplotype is based on the haplotype of the first SNV in the local normal phase block; that haplotype is always arbitrarily assumed to be 1. The normal and tumor haplotypes are then set to be congruent to one another by comparing positions heterozygote in both samples. In case the normal region is not phased, the tumor haplotype is assumed to be 2. If the tumor SNV is a homozygote and the normal is a heterozygote, the haplotype is assigned to the wild-type haplotype of the normal.

Cancer genome allelic imbalance analysis. For assessment of LOH events, our analysis relied on minor allelic frequency (MAF) data. The MAF is a ratio comparison of allelic read depths from heterozygous SNVs identified from the normal genome compared to the same position from the tumor. The input file is a VCF containing the normal and tumor reads. The calls are filtered to require a genotype quality (GQ) of 30 or greater in both normal and tumor tissue at that position, an overall read depth of 10 or greater, and a minor allele depth of at least 3 in the normal genome. The allele depth ratio is calculated as the minor allele count divided by the major allele count. The MAF value is determined as follows: we divide the tumor allele depth ratio by the normal depth ratio and taking the $\log_2$ of the quotient. For graphic display, we used a smoothed MAF value based on a window average of 100 contiguous SNVs from each genome.

Cancer genome CNV and SV analysis. To determine somatic copy number alterations and the affected genomic intervals from WGS data, we used the SeqCBS method[31]. The software implementation is available as an open-source R package named SeqCBS (https://cran.r-project.org/). The CNV analysis used an R script that reads a configuration file listing the sequence data sets to be compared, namely the case (tumor) versus the control (normal). The algorithm then performs the segmentation on these two files, compares them, and produces both local and whole-chromosome CNV plots. For any such region, there is a general test statistic and a relative gain or loss copy number value. Generally, we required a test statistic >1,000 as a basic cutoff and a copy number value of >2.5 or <1.6 as our thresholds for marking an event as a significant amplification or loss. We validated these calls with linked reads by counting the average number of barcodes-annotated reads over a 50-kb window spanning the length of each candidate.

To validate SV calls made by the GemCode software analysis of linked reads we examined sequence data from the short-read WGS data set. We used BreakDancer[30] and its default setting to generate a set of SV candidates and identified putative locations as predicted by the phased SV call set and associated quality score. In addition, we identified soft-clipped reads in the vicinity of the breakpoints, which are indicative of a SV breakpoint. Afterwards, we tabulated the number of reads directly supporting the breakpoint. Soft-clipped reads were manually curated in IGV (https://www.broadinstitute.org/igv/) to verify base quality and were individually aligned in BLAT[40] to verify the breakpoint locations.

40. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

doi:10.1038/nbt.3432